Tabla de Contenidos

Contextos semánticos y análisis argumental como propuesta para mejorar búsquedas en Internet

— *Joaquín Herrero Pintado* *jherrero 2009/11/14 11:54*

Introducción

El gran éxito de Internet como gran repositorio documental de información en distintos formatos (texto, imágenes, vídeo y audio) está reviviendo el debate sobre si el modelo actual de indexación de contenido que llevan a cabo los buscadores es suficiente para localizar la información que un usuario especializado necesita.

La dispersión de la información en Internet junto con el hecho de que cada documento está encerrado en un formato distinto son elementos que dificultan aún más la tarea de trazar el mapa temático de los contenidos de Internet.

El presente trabajo hace una propuesta para iniciar el camino hacia la agrupación temática de los contenidos de Internet que se aprovecha de las inercias naturales de los usuarios en su búsqueda de información para, con el apoyo de las redes sociales y el uso de técnicas de análisis argumental, conseguir crear índices temáticos que permitieran no solo localizar la documentación temáticamente, sino también construir de forma natural buscadores especializados.

Cuestiones sobre el uso de Internet como repositorio documental

Desde la perspectiva que trata el presente trabajo, podemos considerar a Internet como un conjunto geográficamente disperso de textos sin ninguna agrupación temática que los una. Aunque existen diversas webs especializadas en las que podemos encontrar un conjunto de textos temáticos y algún tipo de buscador propio para localizarlos, difícilmente se puede considerar a Internet en su totalidad como una gran enciclopedia que recoja todo el saber, sino más bien como un apilamiento desorganizado pero más o menos usable de documentos que recogen todo el saber humano.

Para poder localizar la documentación que precisamos, disponemos de buscadores como Google, Bing, Yahoo o Ask, que indexan los contenidos de los documentos y nos permiten recuperar los que contienen cierta secuencia de palabras, diferenciándose unos buscadores de otros en el criterio por el que ordenan la información que presentan al usuario.

Para indexar los documentos dispersos en Internet, los servicios de búsqueda disponen de un software (llamado “bot”, robot o “spider”, araña) que simula ser un visitante de un sitio web y lo recorre en su totalidad, descargando el contenido de dichas páginas a una base de datos propia de

cada buscador. Dicha base de datos es indexada en función de distintos criterios según sea el buscador que lo haga.

Sin embargo los buscadores no nacieron con la vocación de recorrer e indexar Internet. Yahoo, por ejemplo, nació como un índice temático, confeccionado manualmente, de las páginas que más le gustaban a Jerry Yang, su creador. Pero el crecimiento exponencial del número de páginas en Internet hizo que se descartara en Yahoo el procedimiento manual de confección del índice y se pasara a la técnica comentada de recorrer Internet e indexar el contenido de las páginas visitadas.

Google se diferenció (y aún lo hace) del resto de buscadores por la forma de indexar las páginas que visita. Su algoritmo, llamado “page rank”, puntúa las páginas en función de su popularidad en Internet, usando ese criterio para priorizar unos resultados sobre otros y presentarlos en primer lugar al usuario que hace una búsqueda. Al presentar en primer lugar las páginas más populares, la mayoría de los usuarios encuentran lo que buscan en menos intentos que en otros buscadores, lo que explica el éxito de este buscador.

Se estima que en Mayo de 2009 hay casi 110 millones de sitios web en Internet, cada uno de ellos con decenas o miles de documentos sobre diferentes temas. El hecho de que los buscadores actuales utilicen la popularidad como único criterio de ordenación de resultados es incompatible con el aumento de contenido especializado en Internet, por lo que resultan muy ineficaces como instrumento de búsqueda especializada.

Se puede decir que los buscadores actuales indexan la dispersión pero no la solucionan. Es un hecho reconocido que existe la necesidad de categorizar y relacionar la información que presentan los buscadores, pero ello nos pone frente al problema de cómo generar automáticamente dichas categorías y relaciones, siendo imposible hacerlo manualmente, como demostró Yahoo hace años. La aparición de la llamada “Web 2.0” y más concretamente del concepto de “redes sociales”, una estructura social cuyos nodos (sean personas individuales u organizaciones) se agrupan en función de intereses comunes, ha marcado un antes y un después en el uso que se hace de Internet.

Si las redes sociales son una agrupación de personas en torno a un interés común, cabría caracterizar a las redes sociales en función del tipo de interés común, distinguiéndose así dos tipos de redes sociales: las que tienen como centro la intercomunicación personal (Facebook) o las que agrupan a sus miembros en torno a la información (Wikipedia).

Las redes sociales han resultado muy eficaces en producir “inteligencia colectiva” cuando se han agrupado en torno a la información. Un ejemplo del resultado de dicho esfuerzo conjunto es la propia Wikipedia, de cuyo contenido se calcula que en un 73% ha sido producido por el trabajo en colaboración de unas 1400 personas.

Pero hay otros ejemplos aún más interesantes de los logros de las redes sociales, que tienen relación con lo que aqui exponemos:

- El “social tagging”, también llamado “folcsonomía”, o “indexación social”, es decir, la clasificación colaborativa de la información en internet por medio de asignarle etiquetas descriptivas simples
- El “social bookmarking”, o “marcadores sociales”, que es una forma de almacenar, clasificar y compartir los enlaces a contenido de internet etiquetados en función del tema

Este marcado e indexado social se ha aplicado a todo tipo de contenido en Internet, como fotografías, vídeos y documentos de texto en sus diversos formatos de presentación, por lo que gracias a este

trabajo colaborativo disponemos actualmente de mucha precisión a la hora de tipificar la información de Internet, al haber sido enriquecido el contenido con los términos aportados por el marcado social, términos que pueden ser objeto de búsqueda por parte de los buscadores antes comentados.

El presente trabajo pretende hacer una aproximación a un nuevo uso de las redes sociales en el ámbito de la búsqueda de información en Internet: la creación de contextos semánticos para agrupar temáticamente la información en Internet y la aplicación a dichos contextos de técnicas de análisis argumental.

La propuesta del contexto semántico

¿Qué sería un "contexto semántico" en Internet? Sería un repositorio, almacén o contenedor usado para recopilar ciertos datos acerca de la información dispersada en internet, pero relativa a un solo tema. Por ello también lo podríamos llamar "contenedor temático".

La información allí recolectada sería como mínimo esta:

- la URL (dirección) donde está localizada en Internet el documento original
- el contenido a texto completo de dicho documento
- las descripciones y etiquetas asignadas a dicho documento por parte de las redes de marcadores sociales

Una característica de este contenedor es que no almacenaría la maquetación original de la página (la página en su formato original puede ser consultable en su dirección), pero recolecta dos elementos clave para iniciar un proceso de análisis semántico:

- las descripciones y etiquetas asignadas por redes de marcadores sociales
- su contenido a texto completo

Dado que el contenedor aquí propuesto puede ser considerado desde dos perspectivas diferentes, usaremos indistintamente para referirnos a él los términos *contenedor temático* y *contexto semántico* pues ambos son los usos básico y avanzado de la misma propuesta: con el contenedor los usuarios reciben el servicio de "lista de favoritos" que usan tradicionalmente, y desde ese punto de vista estamos hablando de un "contenedor temático", pero desde el punto de vista de la creación de superestructuras temáticas en Internet el mismo contenedor puede ser visto y tratado como un "contexto semántico" y aplicarle técnicas de análisis argumental.

Uno de los problemas de la información que encontramos en Internet es que no solamente está dispersa, sino además "encarcelada" en diversos formatos que requieren programas específicos para que podamos obtener la información allí contenida. Algunos de estos formatos son los propios (propietarios) de los diversos procesadores de texto, o formatos de presentaciones tipo "power point", o incluso páginas web estáticas, en las que el texto informativo está mezclado con las instrucciones de maquetación, como es el caso de las páginas HTML o XML.

Dentro del contexto semántico la información recolectada sería "liberada" de su formato y aparecería como texto simple, requisito fundamental para efectuar un análisis eficiente de su contenido.

Por poner un ejemplo: un contenedor temático sobre "¿qué es el proceso de Bolonia?" podría contener varias decenas de referencias a diferente contenido en Internet que lo explique, como páginas web o presentaciones "power point", pero además contendría las descripciones y etiquetas

sociales de cada contenido, y su contenido a texto completo. Eso hace posible procesar de la siguiente forma el contenido con vistas a analizarlo semánticamente:

- **Romper** en unidades textuales más pequeñas el contenido de cada documento introducido en el contexto semántico
- **Etiquetar** y relacionar entre sí jerarquicamente las pequeñas unidades textuales para que cada una de ellas constituya una pieza de información susceptible de ser aportada como respuesta a una consulta
- **Analizar** argumentalmente a todas las unidades textuales del contenedor
- **Agrupar** las unidades textuales procedentes de diferentes fuentes en función del resultado del análisis argumental, lo que daría cierta unidad de contenido a la información del contenedor aunque haya procedido de distintas fuentes

Pero, ¿mediante qué procedimiento se agregaría la información dispersada en Internet a un contenedor semántico? ¿Y qué clase de análisis se aplicaría al texto allí almacenado?

El buscador como nudo central de una red de contextos semánticos

En la actualidad, a partir de uno o varios términos de búsqueda que aporta un usuario a un buscador, se le presentan los resultados que el buscador considera más relevantes en función de su popularidad en Internet.

Una vez que el buscador nos presenta la información, lo único que podemos hacer es visitar la página propuesta y, si es de nuestro interés, anotar de alguna forma su dirección para una posterior consulta. Este modelo presenta dos principales problemas: la cantidad de resultados obtenidos y el criterio para juzgar la calidad de los mismos.

Respecto al primero, resulta frustrante que al buscar, por ejemplo usando Google, la palabra “argumentación”, obtengamos más de un millón de resultados, de los cuales tenemos delante los diez primeros. Suponiendo que en visitar cada resultado y decidir si es de nuestro interés empleáramos un minuto, tardaríamos un millón de minutos, es decir, casi 700 días, en revisar todo el contenido de Internet con ese término. Es cierto que podemos aportar más términos y así disminuir el número de resultados. Por ejemplo, para “argumentación falaz” Google indica 100.000 resultados, y para “argumentación falaz política”, 70.000, pero sigue siendo una cantidad de información que está más allá de lo abarcable por cualquier persona que esté investigando un tema, por lo que lo único posible es visitar cada uno de los resultados confiando en que el criterio de popularidad usado por el buscador coincida con el nuestro y seguir visitando resultados hasta que se agote nuestro tiempo o nuestra paciencia.

En segundo lugar, se crea un problema en la atribución de calidad a una página: ¿son los primeros resultados de Google los que tienen más calidad o se les adjudica más calidad porque aparecen como primeros resultados? No es infrecuente escuchar que algunos sitios web muy relevantes para el tema buscado no aparecen en las primeras páginas de resultados porque el diseñador de la página no la ha optimizado suficientemente, o como se suele decir, no ha hecho bien el SEO (Search Engine Optimization). El criterio de bondad de la información contenida en una página no puede depender del diseño estructural de la misma, porque esto crea el problema de que una página tenga estructura

de calidad (meta etiquetas relevantes, términos frecuentes en los títulos, etc.) pero contenga información irrelevante. Es más, existe el peligro de que por acostumbrarnos a visitar únicamente los primeros resultados obtenidos en una búsqueda vayamos modificando progresivamente nuestro criterio personal de calidad para hacerlo coincidir con la información que el buscador nos proporciona, y de esa manera llegamos a atribuir calidad a un resultado que simplemente es popular.

Este mínimo común criterio de calidad que se nos impone desde los buscadores que depende en parte del diseño estructural de la página y del número de enlaces que la apunten implica que quedan anulados los criterios de calidad de los expertos. Una búsqueda en Internet que pretendiera ser de calidad debería de incluir lo que podríamos llamar “resultados de autor”, es decir, enlaces a aquellas páginas que un experto en la materia considera imprescindibles para estar bien informados sobre el tema, independientemente de si son populares o si la estructura de la página contiene las palabras adecuadas en los lugares adecuados.

Sin embargo, es posible hacer de la necesidad virtud y mezclar los dos conceptos que hemos manejado hasta ahora: los actuales buscadores y el modelo de red social cooperativa para empezar a andar hacia el objetivo de la agrupación temática del contenido en Internet.

Mediante el gesto estandarizado en la informática personal de “pinchar y arrastrar”, debería de ser posible “echar” cualquiera de los resultados presentados por el buscador directamente desde la lista de resultados al contenedor temático antes explicado para su posterior revisión por parte del usuario.

Si este sencillo gesto estuviera incorporado en los buscadores se resolverían varios problemas al mismo tiempo:

- Permitiría al usuario tener siempre a mano los resultados que él mismo ha seleccionado con tan solo abrir el contenedor en el que echó los resultados que le interesaron
- Suponiendo que el resultado sea relevante respecto al tema del contenedor, habríamos tematizado el resultado seleccionado y habríamos enriquecido el tema en cuestión

Por tanto la tematización o más bien, la agrupación tematizada del contenido de Internet sucedería de forma gradual y absolutamente natural, reutilizando los mismos hábitos de los usuarios que necesitan agregan un resultado a la lista de favoritos de su navegador o a una red social de marcadores.

La existencia en la actualidad de redes sociales muy populares, como “delicious.com”, cuyo único objetivo es guardar enlaces a contenido web de interés, agregando a cada enlace un comentario y etiquetas temáticas, indica que el camino que propongo hacia el contexto semántico como un contenedor de información de interés ya ha sido iniciado en sus aspectos más básicos.

Si se facilitara la compartición de estos contenedores temáticos con otros usuarios de Internet, dichos contenedores se podrían convertir en puntos focales de pequeñas redes sociales temáticas, lo cual ampliaría el concepto actual de red social, dejando de ser tan solo una tupida malla de relaciones interpersonales como por ejemplo también lo son son las redes que se crean en las reuniones sociales, que no son “alrededor” de nada sino “con” muchos, y pasarían a ser una malla de relaciones interpersonales alrededor de un tema, tal como sucede en las reuniones de trabajo, en las que se crean redes interpersonales condicionadas por la existencia de un tema de la reunión que metafóricamente está representado por la mesa alrededor de la cual están reunidos.

El concepto de red social aplicado a la estructuración del contenido de Internet en contenedores temáticos es lo que haría viable ahora lo que en su día fue imposible para Yahoo: la catalogación por temas de los recursos de internet, debido a que ahora no se necesitaría el esfuerzo deliberado de una

clasificación manual sino que la agrupación temática de la información sucedería como consecuencia natural del uso de Internet.

El análisis argumental dentro del contexto semántico

Hasta ahora hemos visto el contexto semántico como un contenedor que aloja información de manera parecida a como lo hacen en la actualidad las redes de marcado y etiquetado sociales. Sin embargo veremos ahora el contenedor propuesto desde una nueva óptica, la del análisis argumental.

Dado que el contenedor al que hemos llamado “contexto semántico” es relativo a un solo tema, toda la información que contenga puede ser referida a un mismo marco común de supuestos sin que sea necesario construir dicho marco a partir del análisis del contenido de cada documento, sino que sería deducido del hecho de estar dentro de un contenedor que está referido a un tema concreto y que jerárquicamente puede estar asociado con una serie de conceptos semánticos, como veremos más adelante.

El etiquetado social que incluye el contenido que se 'arrojaría' al contenedor podría ser usado para componer una primera versión de un diccionario de términos para uso interno del contexto, diccionario que mediante un tesauro podría incluso estar ordenado conceptualmente.

¿Cómo elaboraríamos el marco común de supuestos que permitiría inferir el significado del contenido alojado en el contenedor temático?

Una primera aproximación podría ser mediante localizar en el contenido textual del contenedor todos los asertos, tratando de clasificarlos mediante un análisis de los indicadores de fuerza que contengan para puntuar su grado de influencia.

La “Gramática de la argumentación” de Vincenzo Lo Cascio [1)] contiene en su capítulo 6 un estudio detallado de los indicadores de fuerza, indicando que “*pueden clasificarse según la función que realizan. Algunos marcan la tesis, otros los datos o los argumentos, otros la regla general y otros la reserva, la fuente, o categorías mayores como la argumentación misma*”.

El desafío a la hora de detectar los indicadores de fuerza en un texto es la posibilidad de que, por ser la transcripción de una conversación oral, hayan sido sustituidos por la entonación específica propia del lenguaje oral, o incluso, aunque el texto originalmente sea un escrito, se haya usado como indicador de fuerza un orden deliberado de los enunciados, lo cual solo podría ser interpretado acudiendo a un profundo conocimiento del mundo específico que se narra.

Aunque esto es así, no obstante es posible encontrar en los textos marcadores que pueden usarse con un grado elevado de fiabilidad para calificar una declaración como un dato, una justificación, una opinión o una conclusión, por nombrar algunos.

La clasificación que hace Lo Cascio en las páginas 203 y 204 de los indicadores de fuerza que propongo localizar dentro del contenido del contexto semántico son:

Indicadores de fuerza que introducen un macroargumento	(ahora me explico, el razonamiento es éste, ahora se demuestra por qué)
Introducen un argumento o un dato: JUSTIFICADORES	(puesto que, porque, de hecho, en efecto, dado que, ya que, ya que es cierto que, también porque, considerando que, partiendo del hecho que, y la prueba es que, y eso es porque, luego, uso del gerundio)
Introducen la tesis o conclusión (de primer o segundo nivel): CONCLUSIVOS	(por consiguiente, así pues, por tanto, he aquí que, por eso, se sigue que, por lo cual puede sostenerse que, por ello, si… entonces)
Introducen la regla general: GENERALIZADORES	(a partir de…, dado que…, y eso porque…, dice que…)
Introducen la modalidad o el calificador: MODALES	(quizá, probablemente, es probable que, necesariamente, poder + infinitivo, deber de + infinitivo, futuro [elemento morfológico con funciones modal y no de tiempo verbal])
Introducen la fuente, la autoridad: GARANTES	(como dice, según…)
Introducen una reserva: RELATIVIZADORES	(a no ser que, salvo que, a menos que, excepto que, si / si no, aunque)
Introducen un refuerzo para la justificación presentada: REFUERZOS	(sin contar con, si se tiene en cuenta el hecho de que, observemos que, no obstante, a pesar de que, si bien, aunque)
Introducen una contraopinión: ALTERNANTES	(sin embargo, no obstante que, a pesar de que)

Estos indicadores de fuerza pueden ser fácilmente localizados mediante el uso en los lenguajes informáticos de programación de expresiones regulares como patrones de búsqueda dentro del contenido textual del contenedor y eso permitiría aislar y extraer el argumento que introducen.

En principio, todos los asertos hechos con indicadores de fuerza conclusivos podrían ser considerados parte del marco común de supuestos del contenedor, pues contendrían afirmaciones que se hacen tras una argumentación.

También se podría someter a un análisis cada uno de los asertos conclusivos para filtrar aquellos que sintácticamente son oraciones atributivas, pues en ellas se tendría como sujeto un concepto que estaría en el diccionario de términos antes explicado, y como predicado el atributo que le caracteriza y que podría usarse para relacionar conceptos entre sí y de esa forma ir creando una red que permitiría un análisis conceptual más avanzado del contenido.

Dichos asertos también podrían ser considerados el resumen de la información del contenedor, lo cual haría muy útil el contenedor contextual como herramienta para resumir de forma rápida una gran cantidad de información: se “echan” al contenedor los documentos arrastrando los enlaces desde el buscador donde los hemos localizado, y a continuación el contenedor podría elaborar un informe con todas las conclusiones contenidas en los documentos que se le ha proporcionado.

Por lo ya descrito, el contenedor iría creando en su interior una base de datos en la que organizaría los resultados del análisis al que se va sometiendo a la información, la cual contendría:

- Los asertos con indicadores de fuerza concluyentes
- La red-diccionario de conceptos
- La red-diccionario de etiquetas sociales
- La red-diccionario de palabras frecuentes

Cada uno de los diccionarios que se detallan (de conceptos, de etiquetas sociales y de palabras frecuentes) debería de poder usarse para constituir una red interna semejante a la red de transportes de una ciudad, de tal modo que yo pueda elegir uno o varios términos de cada uno de los diccionarios y recorrer la información del contenedor que está relacionada con dichos términos.

Estas redes de transporte internas del contenedor son las que harían posible usarle no solo para mantener reunida información que previamente estaba dispersa, sino también para someterle a búsquedas y que pueda actuar así como proveedor de información proporcionando respuestas. Las respuestas obtenidas de tal contenedor especializado serían de mucha más calidad de las obtenidas en un buscador generalista.

El contexto semántico como proveedor de información

Si se hace la integración propuesta entre los buscadores de internet y los contextos semánticos que se pondrían a disposición de los usuarios para su uso personal, el propio uso natural de Internet haría que la información se agrupara por temas.

Tal como en la actualidad hacen algunos buscadores, que consideran la Wikipedia como una fuente de información independiente de criterios de popularidad por lo que proporcionan sus definiciones en primer lugar en su lista de resultados (así hace Google, por ejemplo), podría llegar un momento en el que fuera más rentable en términos de esfuerzo computacional ofrecer al usuario uno o varios contenedores temáticos como resultado de su búsqueda que las decenas de enlaces individuales a la misma información en su versión dispersada en Internet.

En ese escenario, si como resultado de su búsqueda, en vez de obtener una interminable lista de documentos individuales, los buscadores proporcionaran contenedores temáticos, el usuario cada vez que buscara obtendría mucho más que ahora, pues el contenedor le pondría en contacto con:

- La documentación que busca en forma catalogada y ordenada por el grupo de personas que mantiene el contenedor temático que se le presenta
- Contacto con la red social que creó y mantiene la información alojada en dicho contenedor, que podría ser tan pequeña como una sola persona, o tan grande como una universidad que hubiera decidido tematizar la información que produce dentro de contenedores

Redes de contextos semánticos

Los contenedores deberían de tener en su interior no solo documentos, sino los programas que actuaran sobre la información que contienen y que, además, buscarían en Internet a contenedores semejantes para fabricar redes sociales de contenedores.

La forma en que los contenedores localizarían a sus semejantes sería mediante el uso de tesauros, que podrían crear redes de contextos de la siguiente forma:

- Determinando los conceptos de jerarquía superior (hiperónimos) con los que un determinado contenedor está relacionado
- Mediante la adscripción de cada contenedor a uno o varios contextos-hiperónimos
- Localizando sinónimos de los términos del diccionario que definen al contenedor
- Propiciando la agrupación o fusión de contenedores sinónimos entre sí para evitar duplicidades

Los contextos semánticos como fuente de información para buscadores especializados

La agrupación de contextos en jerarquías conceptuales propuesta en el apartado anterior junto con el hecho de que cada red de contenedores dispone de un diccionario de términos relevantes, haría posible la creación de buscadores especializados, que podrían usar la información de los contextos semánticos para elaborar diversas estrategias de búsqueda:

- Si alguien quiere buscar en Internet información sobre un tema del que existe una red de contenedores temáticos, los términos de búsqueda propuestos por el usuario pueden ser completados con los términos procedentes del diccionario de la jerarquía de contenedores, de tal forma que la búsqueda aumenta su concrección y eficacia
- A partir de las direcciones de Internet de las que procede el contenido de los contenedores temáticos es posible elaborar una lista de sitios web de búsqueda prioritaria para cada tema o concepto, con lo que la búsqueda se concentraría preferentemente en los lugares de Internet que más información producen para dicho tema de búsqueda

Conclusiones

La actual configuración de Internet como red de documentación dispersada y encerrada en formatos más o menos propietarios no permite usar la información como conocimiento.

La creación de contenedores temáticos asociados a buscadores y alrededor de los cuales se presten servicios a personas integradas en redes sociales permitiría de forma gradual crear una estructura temática superior a los índices de los actuales buscadores; y la aplicación a dichos contenedores de técnicas de análisis argumental permitiría la extracción de información relevante, que permitiría convertir la red de documentos actual en una red temática de conocimiento.

1)

Gramática de la argumentación, Vincenzo Lo Cascio, Alianza Editorial, Madrid, 1998